Exploring Machine Learning to Study the Long-Term Transformation of News

Broersma, Marcel; Harbers, Frank

Published in:
Digital Journalism

DOI:
10.1080/21670811.2018.1513337

Volume 6 Number 9 November 2018

Routledge
Taylor & Francis Group

DIGITAL JOURNALISM

Digital Journalism

ISSN: 2167-0811 (Print) 2167-082X (Online) Journal homepage: http://www.tandfonline.com/loi/rdij20

Exploring Machine Learning to Study the Long-Term Transformation of News

Marcel Broersma & Frank Harbers

To cite this article: Marcel Broersma & Frank Harbers (2018) Exploring Machine Learning to Study the Long-Term Transformation of News, Digital Journalism, 6:9, 1150-1164, DOI: 10.1080/21670811.2018.1513337

To link to this article: https://doi.org/10.1080/21670811.2018.1513337

Published online: 11 Oct 2018.

EXPLORING MACHINE LEARNING TO STUDY THE LONG-TERM TRANSFORMATION OF NEWS
Digital newspaper archives, journalism history, and algorithmic transparency

Marcel Broersma and **Frank Harbers**

The labour-intensive nature of manual content analysis and the problematic accessibility of source material make quantitative analyses of news content still scarce in journalism history. However, the digitization of newspaper archives now allows for innovative digital methods for systematic longitudinal research beyond the scope of incidental case studies. We argue that supervised machine learning offers promising approaches to analyse abundant source material, ground analyses in big data, and map the structural transformation of journalistic discourse longitudinally. By automatically analysing form and style conventions, that reflect underlying professional norms and practices, the structure of news coverage can be studied more closely. However, automatically classifying latent and period-specific coding categories is highly complex. The structure of digital newspaper archives (e.g. segmentation, OCR) complicates this even more, while machine learning algorithms are often a black box. This paper shows how making classification processes transparent enables journalism scholars to employ these computational methods in a reliable and valid way. We illustrate this by focusing on the issues we encountered with automatically classifying news genres, an illuminating but particularly complex coding category. Ultimately, such an approach could foster a revision of journalism history, particularly the often hypothesized but understudied shift from opinion-based to fact-centred reporting.

KEYWORDS journalism history; machine learning; (automatic) content analysis; digital newspaper archives; digitization; news genre; algorithms; algorithmic transparency

Introduction: From Scarcity to Abundance

Access to old news has been improved tremendously in the past decades. National libraries in for example France, Australia and the Netherlands have digitized their historical newspaper collections on a large scale while many local archives have digitized individual titles that cater to the interests of regional historians. In contrast to this "public model" which provides free access to everyone, there is a "commercial model" that has been applied in countries such as the US and the UK. Here publishers

Digital Journalism, 2018
Vol. 6, No. 9, 1150–1164, https://doi.org/10.1080/21670811.2018.1513337

have sold their rights to commercial companies such as Cengage, which have digitized papers and created databases to sell subscriptions to universities, public libraries and archives (cf. Nicholson 2013; Mussell 2008), hampering the access to research material. Nevertheless, cost considerations and accessibility issues aside, research can now be done from behind the desk and is thus less time-consuming.

Digitization has clearly opened up new venues for journalism research because large text corpora are now full-text searchable. However, the quality and availability of historical newspaper archives diverges considerably, and with that their value for research. First, access to sources differs between countries and between archives. While public archives might under special conditions be inclined to provide researchers access to the complete data set of text files (and metadata), commercial companies are hesitant because it jeopardizes their business model. Second, because of copyright issues periodicals are often not available after 1945. This puts severe restraints on comparative and longitudinal research. Third, the enthusiastic uptake of large-scale digitization projects and the advancement of technology have resulted in public, semi-public and commercial silos using different digitization standards and procedures. This not only prohibits data integration on a higher level, but also results in different quality of text files due to issues with OCR and segmentation. The number of errors in text files generated in the digitization process also differs tremendously between historical periods and publications (cf. Broersma 2011a, 2011b; Wijfjes 2017).

Even more importantly, these archives generally only provide keyword search possibilities, based on simple and more complex queries applying for example (Boolean) operators and wildcards, as the main gateway to their content. Therefore, data can usually only be searched and retrieved through the search interface, resulting in a list of individual hits. While this is a major step forward compared to endlessly scrolling through microfilms or turning pages, it remains unsatisfying. Keyword search only affords straightforward queries, as Deacon (2007, 8) argues: "key word searching is best suited for identifying tangible 'things' (i.e. people, places, events and policies) rather than 'themes' (i.e. more abstract, subtler and multifaceted concepts)." Moreover, metadata, which can be used to limit search results, tend to be added sparsely.

The keyword search tools and the opportunity to download pages in PDF-format most archives offer, are usually sufficient to cater to the demands of the general audience and scholars who consult historical newspapers as a source of specific historical information about a certain event or issue. However, it limits more specialized data publics such as journalism scholars and media historians, who want to study newspapers as a serial source and are interested in the structural transformation of news and journalism. The key question is, therefore, whether and how digitization will actually change research practices in journalism history (cf. Boumans and Trilling 2016; Flaounas et al. 2013). Despite the development of (computer-assisted) social scientific ways of research such as (automatic) quantitative content analysis that offer the opportunity to explore news content beyond ideographic and myopic studies, journalism historians have been reluctant in adopting quantitative and computational methods (Wijfjes 2017; Nicholson 2013; Broersma 2011a, 2011b).

We join in calls for journalism scholars to move beyond keyword search and manual content analysis and take full advantage of the available digitized newspaper material (Boumans and Trilling 2016; Flaounas et al. 2013; Günther and Quandt 2016;

Jacobi, Van Atteveldt, and Welbers 2016; Burscher, Vliegenthart, and de Vreese 2015). Computational methods based on machine learning enable us to root analyses in large data sets instead of necessarily modest samples (Boumans and Trilling 2016; Broersma 2011a, 2011b; Wijfjes 2017). This implies that "we no longer have to choose between data size and data depth" (Manovich 2012, 466). Such approaches not only allow us to ask new questions, but also to come to new conclusions—and challenge the ones not rooted in textual analysis or just based on small subsets of newspaper content. Automatic content analysis, e.g. text statistics, sentiment analysis, topic modelling or frame analysis, facilitates detailed analyses of newspapers as a serial source on an unprecedented scale in a much more cost-efficient way (cf. Boumans and Trilling 2016).

Successfully implementing advanced digital methods could help to systematically study more conceptual questions such as the historical shift of topics and concepts. In addition, we argue for taking a next step by automating content analysis of the formal structure of texts and images. Enabling and facilitating such longitudinal analyses would address an important and peculiar gap in journalism history. Until now, due to methodological and practical issues, research has spent only limited attention to content analysis of the formal characteristics of news (cf. Bingham 2010). The time-intensive nature of this kind of research and the accessibility of analogue newspaper collections put up too high barriers. Though understandable from a practical perspective, this neglect is nevertheless remarkable since the value of journalism for society is first and foremost based on its capacity to provide legitimate representations of social reality, for which form is a crucial category as Broersma (2011b) has argued.

Only recently, scholars, often in interdisciplinary teams of historians, programmers and data scientists, have started to tap into the vast collections of digitized historical news texts. We agree with the growing body of literature on computational methods in journalism research that forms of automated content analysis, specifically machine learning approaches, offer promising venues to analyse big data sets of news content and introduce new questions and approaches to journalism studies (Boumans and Trilling 2016; Flaounas et al. 2013; Günther and Quandt 2016; Jacobi, van Atteveldt, and Welbers 2016; Burschers, Vliegenthart, and de Vreese 2015). It allows for grounding analyses in big data and mapping the structural transformation of journalistic discourse on a large scale.

Still, current approaches mostly focus on recent rather than historical news texts. Digital born data sets are easier to gather, contain less to zero (OCR) errors, and do not have to account for change over time. Furthermore, the emphasis tends to be on the topical content of news texts. The frequency, variety or co-occurrence of words are used as manifest indicators of topics, frames or sentiments in a "stable" synchronic data set (see for instance, Boumans and Trilling 2016; Flaounas et al. 2013; Günther and Quandt 2016; Burscher et al. 2014). While important, this does not provide insights into the *structural transformation* of news and journalism, which sheds light on how journalism "works" beyond day-to-day news stories. While the content of news changes every day, form and style are more stable categories. They indicate professional norms about how journalism needs to be performed and what accounts for a truthful and trustworthy representation of reality. The next step is thus to further develop methods and tools that allow us to analyse historical, diachronic data sets to map the development of news and journalism.

In this article, we argue that a focus on genre, as a marker of professional ideology, enables us to gain important new insights into the development of journalism. Textual characteristics relating to journalistic form conventions and modes of expression, such as genre, are until now rarely discussed or problematized thoroughly in scholarship (cf. Boumans and Trilling 2016). Moreover, we discuss why supervised machine learning is a fruitful and promising approach to automating genre classification and outline how we have operationalized this for our research. We also discuss the issues that complicate the automation of such a complex latent content category. Pivotal in this discussion is the importance of creating a transparent assessment of the performance of machine learning algorithms—too often an opaque black box process. This discussion reveals how in these highly complicated machine-learning tasks transparency is imperative in moving historical research forward.

The Neglect of Newspaper Form in Journalism History

In historical research, most studies that use historical newspaper material do not study or critically reflect on the medium itself and its development. Newspapers are still mainly used to get factual information about historical figures, events and issues—often only to add flavour with telling citations. The medium-specific qualities and its consequences for the source material are often left out of the equation. In this sense not much has changed since historian Brian Maidment argued in 1990 that scientific progress could only be made "if we regard periodicals not like fossil hunters, in search of specimens to fill a cabinet, but like theoretical geologists or theologians, as expositions of processes by which change occurs and is made legible" (quoted in: Vella 2009, 205).

With notable exceptions (Barnhurst and Nerone 2001; Barnhurst 2016; Fink and Schudson 2014), quantitative analyses of news content and textual conventions, inspired by research approaches in the social sciences and aimed at theory building by tracing patterns, remain very scarce in journalism history (see for more details: Broersma 2011a). The history of journalism has largely been studied through archival research into journalism's institutional development and the analysis of discourses *on* journalism, such as public statements, debates and autobiographical writing of journalists. These are taken at face value rather than that the strategic nature of such discourses is critically studied (Broersma 2010a).

This has resulted in a distorted picture of the historical development of journalism (Nerone 2010; Harbers 2014; Broersma 2018). Historians have created a transnational grand narrative, which has become known as the "liberal narrative" (Curran 2009). It frames the development of journalism since the nineteenth century as a linear development from a partisan press to an independent and autonomous press. It emphasizes "the establishment of an autonomous profession that, independent from political and economic powers, obeys more or less to the objectivity regime, and the practices and formal conventions resulting from it" (Broersma 2018; cf. Broersma 2007, 2010b; Harbers 2014). However, longitudinal content analysis suggests that this dominant narrative is actually skewed and overemphasizes the innovative nature and pace of journalistic development (Harbers 2014). Such research reveals what Dahlgren (1992, 7) called "the gap between the realities of journalism and its official presentation of self."

In addition to qualitative textual analysis, quantitative content analysis that traces the development of the textual characteristics of newspaper content can add nuance and complexity to our picture of journalism history. Analysing a representative sample of daily news coverage can elucidate how journalism's modes of expression were employed in everyday practice as well as their historical transformation. It abstracts from the specific content of a news item to broader predefined categories that are traced over time. Specifically, a focus on formal conventions—how a news item is structured and written, and how it is presented to readers—allows us to move beyond day-to-day news events and tap into the underlying structure of news coverage (Broersma 2010b).

These textual conventions pertaining to journalism's form, i.e. the arrangement of layout, genre, and narrative structure and devices, allude to journalism's professional norms and broader cultural discourses (Broersma 2007, 2010b). Such textual conventions can thus reveal "the way[s] the medium imagines itself to be and to act. In its physical arrangement, structure, and format, a newspaper reiterates an ideal for itself" (Barnhurst and Nerone 2001, 3). Studying these formal characteristics historically sheds light on how journalism's modes of expression have gradually transformed and how professional norms and practices, such as the objectivity regime, have emerged and evolved (Broersma 2010a; Benson 2005).

Genre as Marker of Journalistic Style

For our computational approach to the long-term transformation of journalism, we focus on genre as a marker for professional ideology. Genre is an important characteristic of the form of news content. It structures news discourse and signals to readers what they can expect of an article. Specific genres have been invented throughout journalism history to contain new modes of reporting, reflecting the underlying professional ideology. For example, the report, a prolific genre throughout the 19th century that registered meetings and events chronologically and almost verbatim, was replaced in the 20th century by genres such as the reportage, features and the interview that reflect active reporting and highlight the autonomy of the reporter as interpreter of events (Broersma 2007, 2008, 2010b).

Hartsock (2000) makes a useful distinction here between "topical genres" and "modal genres." This reflects an important difference between Anglo-American and European genre conventions. Within the first cultural context genre refers to a *practice* focused around a certain *topic* or "beat." From this perspective any journalistic texts focused on, for example, sport belongs to the genre "sports journalism." In contrast, modal genres—which we study—refer to a set of *formal conventions*, i.e. particular ways of *structuring* texts, which cut across topics. A news article, for example, would be considered a genre. In this case the inverted pyramid model can be regarded as a typical characteristic of how a news report is structured and thus helps to identify this particular genre. An interview is a genre that centres on a conversation between two people and can be discerned by the way the text is structured around questions and answers. In European journalism, such modal genres are considered the cornerstone of the profession. They are central to the training of aspiring journalists and dealt extensively with in textbooks and journalism programs. Students are trained to know the

differences between genres as these are typically used when assigning stories in newsrooms (Broersma 2008).

We define genre as "language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms" (Bhatia 2004 cited in Handford 2010, 258). As such, the rise and use of particular modal genres indicate the identity of newspapers as they refer to certain styles of journalism. By studying genre conventions, we can, therefore, gain insight into the underlying discursive context: the "communicative goals" of journalism and the professional norms and practices they see as necessary to achieve those goals (Broersma 2010b). The interview and the reportage, for instance, only emerged and became popular from the 1880s onwards and are clearly linked to the shift from reflective, opinion-oriented journalism to fact-based and event-centred reporting (Broersma 2008; Harbers 2014). Examining to what extent these genres are used in newspapers and tracing this over time, therefore, offers important insights into the way journalism was conceived and practiced and how it developed historically (Broersma 2007, 2010b; Harbers 2014).

From Manual Content Analysis to a Machine Learning Approach

In our large-scale research project "Reporting at the Boundaries of the Public Sphere. Form, Style and Strategy of European Journalism, 1885–2005" manual quantitative content analysis offered valuable insights.[1] To study the long-term transformation of news we conducted a longitudinal content analysis of nine European newspapers, which resulted in a database with the metadata, i.e. the topic and genre label, amount and type of sources and images and quotation patterns of 125.000 articles. Yet, this is a highly time consuming and, therefore, expensive endeavour. Moreover, the size of the material that can be annotated is still only a small percentage of the total amount of available material (Harbers 2014).

In line with the potential for journalism studies as outlined by Boumans and Trilling (2016), we see automatic content analysis as highly suitable for longitudinal and comparative research into the historical development of newspapers. However, so far it has mostly focused on text mining and topic modelling (Lee and Myaeng 2002). Studying what kind of topics newspapers reported on in certain periods is certainly fruitful as it can reveal the commercial and ideological strategies of publishers who cater to the demands of different audiences, or show how news values have changed over time (Yang, Torget, and Mihalcea 2011). Nevertheless, it does not offer information on how these topics are presented to the readership or if articles adhere to a fact-based or opinion-oriented style of journalism. The formal characteristics of newspaper texts, such as genre, are still largely left aside. To get a more fine-grained analysis of the underlying journalistic conceptions and modes of expression that newspapers have employed throughout history, it is necessary to move to latent categories of analysis relating to form and style conventions.

Genres, and form conventions in general, are complex latent variables with many and complex sub-categories that are hard to operationalize. Classification of such variables requires much interpretation because they relate to concepts that cannot be observed at the surface of the text, "but can be represented or measured by one or

more [. . . .] indicators" (Hair et al. 2010 cited in Neuendorf 2002, 23). Human coders, therefore, need extensive training to reach acceptable inter coder reliability levels (Neuendorf 2002; Harbers 2014). Doing this automatically proves to be an even more difficult task for computers. As Manovich (2015, 22) explains, forms of automatic content analysis have to deal with a "semantic gap." Machines only recognize manifest "low level information." In the case of newspaper articles, this means semantic and lexical characteristics as well as punctuation marks—though the latter are generally excluded during the preprocessing phase (Günther and Quandt 2016).

Texts are often represented as "bags of words," showing the frequency of each word but disregarding their order. Lexical features are mostly used to disregard less interesting types of words, such as prepositions, articles, and adverbs, and zoom in on (proper) nouns (Jacobi, van Atteveldt, and Welbers 2016). Representing the text through these characteristics is very suitable for determining its topic or frame as it displays the semantic particularities of a text directly related to its meaning (Günther and Quandt 2016; Burscher et al. 2014). However, this approach is too limited for classifying formal characteristics such as genre as it does not include textual characteristics like quotes, metaphors or narrative structure and perspective that allude to the form of texts. These are much harder to operationalize in such a way that a machine can distinguish them.

This makes automating research into formal characteristics of texts such as genre an extremely complicated task. Genres cut across topics and, therefore, cannot be discerned based on semantic characteristics—or at least not solely. Take for instance a reportage. This genre not only provides factual information, but also conveys the atmosphere and the experience of witnessing a certain event or issue, which can range from politics to war, sports and lifestyle. Articles are often structured chronologically, depict a detailed picture of the space and surroundings, convey what the reporter and her sources saw and felt, and use imagery to make the experience of "being there" tangible. In addition, contextual indicators such as the article's length, self-classification and position within the newspaper provide cues.

Such characteristics need much human interpretation and are hard to translate to the low level textual characteristics a computer algorithm can deal with (cf. Manovich 2015; Boumans and Trilling 2016). To complicate things even more, genres are ideal-typical discursive constructs. This means that textual manifestations do not always match the characteristics of these constructs perfectly. Deciding when articles that only partially comply with the textual characteristics of a certain genre can still be considered representative of that genre is challenging. In addition, articles might share characteristics with other genres and genres are dynamic constructions that change or fade away over time while new ones emerge. This makes it hard to rigidly define genres as is necessary for quantitative and deductive forms of (automatic) content analysis (Harbers 2014). To deal with the historical variety of texts within different genres two approaches can be taken. Ideally, an algorithm would be trained on the basis of a training set that covers the entire historical period and recognizes the different genres across history. However, at the moment, this might still result in a low accuracy level and an alternative approach is to train different algorithms based on different training sets for different historical periods.

As such, classifying genres of historical newspaper articles with appropriate levels of validity is clearly challenging. However, we argue that (supervised) machine learning offers a promising approach compared to more traditional ways of automatic content analysis. The latter are closely related to the rule-based approach of manual content analysis in which texts are classified according to predefined categories (Lewis, Zemith, and Hermida 2013; Apté, Damerau, and Weiss 1994). Dictionary- and rule-based automatic content analysis operates in a similar way: the machine assigns a label to a text based on the presence of certain textual characteristics, based on lists of specific words or combinations of words relating to a particular topic, theme or concept. This is a rigid and static way of assigning texts to categories (Günther and Quandt 2016; Zamith and Lewis 2015). Moreover, it is very hard to create and validate exhaustive lists as "most people do not know the complete set of words that indicate a particular content category and/or all ways such words can be used" (Burscher 2016, 21).

While a dictionary approach works for certain tasks (such as determining named entities like sources, or the sentiment of an article), it is problematic for automatic classification of formal characteristics of texts. Here, creating a list of particular words is unlikely to work since these characteristics are independent of the content of an article. Supervised Machine Learning (SML) takes a much more open and dynamic approach because the decision process is not predefined. It uses a manually annotated data set based on predetermined coding categories as initial input. This annotated data set is used to develop and train a self-learning algorithm that creates its own discriminatory model to predict which category is the most likely match.[2] A subset of the training data is used to formally evaluate the performance of the algorithm. As such, it also validates the model it uses to predict the genre of a text. After the algorithm is trained and has reached a satisfactory accuracy level, it can be used to classify new texts. In theory, this makes the need for sampling redundant as the entire corpus could be analyzed (cf. Günther and Quandt 2016; Boumans and Trilling 2016; Grimmer and Stewart 2013; Burscher 2016).

Applying Machine Learning to Historical Newspaper Archives

For this kind of research expertise from various disciplines is imperative. In two research projects, we as domain specialists in journalism history, therefore, work closely with collection specialists from different archival institutions, and data and computer scientists.[3] A main challenge is to translate research questions about journalism history to computer science approaches in machine learning. Not only is the past a foreign country, but those who try to map and analyse it also speak different languages. We build on our experiences with these ongoing projects to discuss the opportunities, pitfalls and problems by applying supervised machine learning to automatic genre classification. In addition, we argue for the importance of algorithmic transparency; scholars without computer science expertise should be able to assess the performance of algorithms beyond mere accuracy percentages.

Underlying our research is a manually annotated data set that has been developed in our previous research project into the transformation of European journalism between 1885 and 2005. Although it also contains metadata about French and British newspapers, we only used the genre classifications of a large sample ($N = 33.000$) of

Dutch historical newspaper articles, as we only had access to the Dutch corresponding digitized newspaper content.[4] This subset is used to train and evaluate different of-the-shelf machine learning algorithms to see which performs best. Ultimately, this allows us to make an informed choice about which algorithm is most suitable for doing a specific machine learning task.

Other than a dictionary-based or a traditional rule-based approach, machine learning algorithms can deal with and combine many different features of texts and independently decide which ones are relevant to base classifications on. In our project we trained several supervised machine learning methods, such as Support Vector Machines, Naive Bayes, and Random Forests, to evaluate and compare their performance. While in our manual content analysis human coders reached an accuracy score of 85 per cent for coding genre with corresponding Krippendorf's alpha of 0.83 (Harbers 2014), the automatic classifiers assign the right genre in between 41 and 70 per cent of the cases.[5] Although the best performing algorithms provide promising scores given the very complex task of genre classification, it leaves the reliability of these tools still under par. They need to perform at least above the lowest accepted accuracy score for human intercoder reliability (Kα 0.67) to draw robust conclusions (Riffe, Lacy, and Fico 2005). That being said, new experiments and adding more training data are likely to improve the reliability of automatic genre classification consistently (see Bilgin et al. 2018, for a more elaborate discussion of our approach, experiments and evaluation of different machine learning algorithms).

However, a major issue with only assessing these overall accuracy scores is that they are the result of a black box process that obscures the built-in choices and biases that result from the training of an algorithm. Without insight into how an algorithm operates and assigns a genre label, it is impossible to evaluate its validity. This is important because the algorithm that performs best in terms of accuracy does not necessarily label texts in a valid way. Therefore, it is crucial to compare different algorithms, elucidating their built-in choices and biases in a way that makes the strengths and weaknesses of their performance transparent. In our collaborative project, we are, therefore, developing a virtual workspace—a dashboard—in which researchers can do experiments with different algorithms on the same training set. This dashboard enables scholars to explore and compare the performance of algorithms and test the influence of different discriminatory features. Through different visualization techniques available on the dashboard, this approach elucidates how these algorithms function in a way that is comprehensible for end users with a humanities and social science background and allows for an informed evaluation of the best functioning algorithm for a specific goal.

One particular issue we encountered is the skewed distribution of genre categories in newspapers—and, therefore, also in the training set. Some genre categories, particularly news reports, are disproportionally present in newspaper content, whereas genres such as the interview or the reportage appear less frequently. While it is advised to train algorithms based on a representative training set, this approach can lead to "overfitting" or "overtraining." The algorithm is then likely to show a preference for genres that are overrepresented in the training data. As such, the average performance might look fine, but the algorithm is likely to underperform in identifying underrepresented genres (cf. Zheng, Wu, and Srihari 2004). In selecting a particular machine

learning algorithm, it is, therefore, adamant to be able to evaluate the performance for each genre category. Our comparison showed that the algorithm with the highest overall accuracy score differed considerably in its performance per genre category, e.g. almost perfect on classifying news reports, but very poorly on classifying a reportage. Based on this particular bias, it is a less likely choice for the task it needs to do.

Another issue relates to the specific features of the discriminatory model algorithms use to classify the genre of a text. As we are interested in modal genres and, therefore, the particular way information about a certain topic is communicated, we consider it crucial that the algorithms base their classifications on "modal cues" rather than "topical cues." For this we used "feature importance ranking plots" to see which textual characteristics were deemed relevant in classifying a certain genre. This showed that although some of the algorithms did indeed use modal cues to assign certain genres, such as the reportage, the algorithms displaying the highest level of overall accuracy also show the highest tendency to base their classifications on topical cues. This conflicts with the conceptualization of genre as a category that cuts across topical boundaries.

Conclusion

Grounded in our experiences with both manual content analysis and supervised machine learning approaches to automatic content analysis, we have argued that the latter offers promising venues for exploring journalism history. We have discussed how such computational methods promise to enable the exploration of large data sets without compromising the depth of the analysis, thus allowing for fine-grained comparative and longitudinal research into the history of news and journalism. We propose to move beyond the relatively easy automation of manifest coding categories (e.g. counting numbers of articles containing certain keywords, or word frequencies within articles) and automatic content analysis based on topic modelling (including frame analysis), because latent content categories pertaining to form, such as genre, can provide us with a more fundamental and detailed insight into the structural transformation of journalism. In our focus on a very complex machine learning task, classifying the latent variable of genre, we compare and evaluate the overall performance of different classifiers, while at the same time rendering the bias and operations of algorithms transparent. Such an endeavor reveals not only the opportunities of our computational approach, but also the persistent problems that need to be solved to fully exploit the research possibilities that digital newspaper archives offer.

At the moment, important progress is made in creating more sophisticated ways to automate the time-consuming method of content analysis, specifically using (supervised) machine learning. To reach valid results, this requires a manually constructed training set with clearly defined categories that can be translated to machine-readable textual features, because, as Simon (2001, 87) argued, "the computer is simply unable to understand human language in all its richness, complexity, and subtlety as can a human coder." However, apart from practical issues concerning the structure and quality of the digital archive, before the potential of a supervised machine learning approach can become reality, important questions about the most suitable procedures to train an algorithm, which algorithm performs best, what the biases are of different

possible algorithms, need to be addressed. Only through enhancing transparency about the training process and performance of such algorithms, we can move toward a trustworthy and reliable approach of analysing the formal characteristics of historical newspaper material.

We consider these approaches to automate forms of content analysis as important additions to the research toolbox of media historians. These will be particularly helpful to test hypotheses that have been formulated based on qualitative research on a larger scale. Similarly, they can be used to map broader developments and trace patterns in historical development that can be contextualized, fleshed out and elaborated by close reading and archival research. In our manual content analysis we could demonstrate how the shift in the course of the twentieth century from opinion-based to fact-centred reporting was far more gradual and messy than is often argued in historical scholarship. Our automated analysis can confirm and refine this conclusion based on a far bigger data set that includes more newspapers while also being distributed more equally over time. Combining computational methods with more contextual qualitative approaches is crucial here because extensive knowledge of the relevant media historical context is pivotal to provide sound and meaningful interpretations of the data generated by algorithms. As Lewis, Zamith, and Hermida (2013, 48) argue: "In the allure of computational methods, researchers must not lose sight of the unique role of humans in the content analysis process. This is particularly true of their ability to bring contextual sensitivity to the content, its manifest and latent characteristics, and its place within the larger media ecology."

In the end, it is important to recognize that automating genre classification, and forms of content analysis in general, should not be regarded as the final solution to the issues with conducting this type of research. As Boyd and Crawford (2012, 671) caution, "context is hard to interpret at scale and even harder to maintain when data are reduced to fit into a model." Despite its potential, this approach—and similar approaches—will not replace traditional forms of media historical scholarship, but much rather complement them.

NOTES

1. This NWO-VIDI project (PI: Broersma) was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 276-45-002. For more details, see the PhD thesis of Harbers who conducted one of the subprojects (Harbers 2014).
2. As we are interested in assigning predefined genre labels that mirror the genres that were current in journalism history we leave forms of unsupervised machine learning aside. For an overview of the opportunities and benefits of unsupervised machine learning for journalism studies, see for instance: Boumans and Trilling 2016; Jacobi, Van Atteveldt, and Welbers, 2016.
3. The first project, "Discerning Journalistic Styles," was conducted by Harbers during a digital humanities fellowship at the Dutch National Library (KB) in 2016 and was a first pilot research that, in collaboration with a data scientist, explored ways to automate genre classification of historical newspaper articles: http://lab.

kb.nl/tool/genre-classifier#introduction The second project, "News Genres: Advancing Media History by Transparent Automatic Genre Classification (NEWSGAC; PI: Broersma, Co-applicant: Harbers)" is a bigger follow-up project, funded by CLARIAH/NWO and the Netherlands e-Science Center under project number ADAH.2016.020. It is a collaboration between the Centre for Media and Journalism Studies of the University of Groningen, the national research institute for mathematics and computer science in the Netherlands CWI, the National Library of the Netherlands and the Netherlands Institute for Sound and Vision: https://www.esciencecenter.nl/project/newsgac.

4. The National Library of the Netherlands graciously granted us access to their dataset of digitized newspapers that is the result of a large-scale newspaper digitization program running since 2006.
5. The default accuracy by predicting the majority class or genre (news report) is 46%.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

FUNDING

This work was supported by CLARIAH/NWO and the Netherlands e-Science Center [Grant number: ADAH.2016.020].

REFERENCES

Allen, Robert B., Ilya Waldstein, and Weizhong Zhu. 2008. "Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres." In *Digital Libraries: Universal and Ubiquitous Access to Information*, edited by George Buchanan, Masood Masoodian, and Sally Jo Cunningham, 379–386. New York: Springer.

Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. 1994. "Automated Learning of Decision Rules for Text Categorization." *ACM Transactions on Information Systems* 12 (3): 233–251.

Barnhurst, Kevin. 2016. *Mr. Pulitzer and the Spider: Modern News from Realism to the Digital*. Urbana: University of Illinois.

Barnhurst, Kevin, and John Nerone. 2001. *The Form of News. A History*. New York: Guildford Press.

Benson, Rodney. 2005. "Mapping Field Variation: Journalism in France and the United States." In *Bourdieu and the Journalistic Field*, edited by Rodney Benson and Eric Neveu, 85–112. Cambridge: Polity Press.

Bilgin, Aysenur, et al. 2018. "Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History." Paper Submitted to the 14th IEEE eScience Conference, Amsterdam. [Under review]

Bingham, Adrian. 2010. "The Digitization of Newspaper Archives: Opportunities and Challenges for Historians." *Twentieth Century British History* 21 (2): 225–231.

Boumans, Jelle W., and Damian Trilling. 2016. "Taking Stock of the Toolkit." *Digital Journalism* 4 (1): 8–23.

Boyd, Danah, and Kate Crawford. 2012. "Critical questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–679.

Broersma, Marcel. 2007. "Form, Style and Journalistic Strategies." In *Form and Style in Journalism. European Newspapers and the Representation of News. 1880-2005*, edited by Marcel Broersma, ix–xxix. Leuven: Peeters.

Broersma, Marcel. 2008. "The Discursive Strategy of a Subversive Genre." In *Vision in Text and Image: The Cultural Turn in the Study of Arts*, edited by Mary Kemperink and Herman, 143–158. Leuven: Peeters.

Broersma, Marcel. 2010a. "De Transformatie van het Journalistieke Veld: Discursieve Strategieën en Journalistiek Vormen." *Tijdschrift voor Communicatiewetenschap* 38 (3): 267–275.

Broersma, Marcel. 2010b. "Journalism as a Performative Discourse. The Importance of Form and Style in Journalism." In *Journalism and Meaning-Making: Reading the Newspaper*, edited by Verica Rupar, 15–35. Cresskill: Hampton Press.

Broersma, Marcel. 2011a. "Nooit Meer Bladeren. Digitale Krantenarchieven als Bron." *Tijdschrift voor Mediageschiedenis* 14 (2): 29–55.

Broersma, Marcel. 2011b. "From Press History to the History of Journalism. National and transnational features of Dutch scholarship." *Medien & Zeit* 26 (3): 17–28.

Broersma, Marcel. 2018. "Americanization, or: The Rhetoric of Modernity. How European Journalism Adapted US Norms, Practices and Conventions." In *The Handbook of European Communication History*, edited by Klaus Arnold, Paschal Preston, and Susanne Kinnebrock. Chichester and Malden: Wiley.

Burscher, Björn, Daan Odijk, Rens Vliegenthart, Maarten de Rijke, and Claes H. de Vreese. 2014. "Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis." *Communication Methods and Measures* 8 (3): 190–206.

Burscher, Björn. 2016. "Machine Learning-Based Content Analysis: Automating the Analysis of Frames and Agendas in Political Communication Research." PhD diss., University of Amsterdam.

Burscher, Björn, Rens Vliegenthart, and Claes H. de Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *Annals AAPS* 659: 122–131.

Curran, James. 2009. "Narratives of Media History Revisited." In *Narrating Media History*, edited by Michael Bailey, 1–21. London: Routledge.

Dahlgren, Peter. 1992. "Introduction." In *Journalism and Popular Culture*, edited by Peter Dahlgren and Colin Sparks, 1–23. London: Sage Publications.

Deacon, David. 2007. "Yesterday's Papers and Today's Technology. Digital Newspaper Archives and "Push Button" Content Analysis." *European Journal of Communication* 22 (1): 5–25.

Fink, Katherine, and Michael Schudson. 2014. The Rise of Contextual Journalism, 1950s–2000s." *Journalism* 15 (1): 3–20.

Flaounas, Ilias, et al. 2014. "Research Methods in the Age of Digital Journalism." *Digital Journalism* 1 (1): 102–116.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Grosan, Crina, and Ajith Abraham. 2011. *Intelligent Systems*. Berlin: Springer.

Günther, Elisabeth, and Thorsten Quandt. 2016. "Word Counts and Topic Models." *Digital Journalism* 4 (1): 75–88.

Handford, Michael. 2010. "What Can a Corpus Tell Us About Specialist Genres." In *The Routledge Handbook for Corpus Linguistics*, edited by Anne O' Keeffe and Michael McCarthy, 255–269. New York: Routledge.

Harbers, Frank. 2014. "Between Personal Experience and Detached Information. The Development of Reporting and the Reportage in Great Britain, the Netherlands and France, 1880-2005." PhD diss., University of Groningen.

Hartsock, John. 2000. *A History of American Literary Journalism. The Emergence of a Modern Narrative Form*. Amherst: University of Massachusetts Press.

Jacobi, Carina, van Atteveldt Wouter, and Kasper Welbers. 2016. "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." *Digital Journalism* 4 (1): 89–106.

Lee, Yong-Bae, and Sung H. Myaeng. 2002. "Text Genre Classification with Genre-Revealing and Subject-Revealing Features." *SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. New York: ACM. 145–150.

Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. 2013. "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods." *Journal of Broadcasting & Electronic Media* 57 (1): 34–52.

Manovich, Lev. 2012. "Trending: The Promises and the Challenges of Big Social Data." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 460–475. Minneapolis, MN: University of Minnesota Press.

Manovich, Lev. 2015. "Data Science and Digital Art History." *International Journal for Digital Art History* 26 (1): 11–35.

Mussell, James. 2008. "Ownership, Institutions, and Methodology." *Journal of Victorian Culture* 13 (1): 94–100.

Nerone, John. 2010. "Genres of Journalism History." *The Communication Review* 13 (1): 15–26.

Neuendorf, Kimberley A. 2002. *The Content Analysis Guidebook*. Thousand Oaks: Sage.

Nicholson, Bob. 2013. "The Digital Turn." *Media History* 19 (1): 59–73.

Riffe, Daniel, Stephen Lacy, and Frederick Fico. 2005. *Analyzing Media Messages. Using Quantitative Content Analysis in Research*. Mahwah: Lawrence Erlbaum Associates.

Simon, A. F. 2001. "A Unified Method for Analyzing Media Framing." In *Communications in U.S. elections: New agendas*, edited by R. P. Hart and D. R. Shaw, 75–89. Lanham, MD: Rowman and Littlefield.

Vella, Stephen. 2009. "Newspapers." In *Reading Primary Sources. The Interpretation of Texts from Nineteenth- and Twentieth-Century History*, edited by Miriam Dobson and Benjamin Ziemann, 192–208. London and New York: Routledge.

Wijfjes, Huub. 2017. "Digital Humanities and Media History. A Challenge for Historical Newspaper Research." *Tijdschrift voor Mediageschiedenis* 20 (1): 4–24.

Yang, Tze-I., Andrew J. Torget, and Rada Mihalcea. 2011. "Topic Modeling on Historical Newspapers." In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 96–104.

Zheng, Zhaohui, Xiaoyun Wu, and Rohini Srihari. 2004. "Feature Selection for Text Categorization on Imbalanced Data." *ACM Sigkdd Explorations Newsletter* 6 (1): 80–89.

Marcel Broersma (author to whom correspondence should be addressed), Centre for Media and Journalism Studies, University of Groningen, The Netherlands. E-mail: m.j.broersma@rug.nl. ORCID http://orcid.org/0000-0002-7342-3472

Frank Harbers, Centre for Media and Journalism Studies, University of Groningen, The Netherlands. E-mail: f.harbers@rug.nl. ORCID http://orcid.org/0000-0003-1578-7582